

# DEVIATION MEASURES IN GENERALIZED LINEAR REGRESSION

*R. Tyrrell Rockafellar*<sup>1</sup>, *Stanislav Uryasev*<sup>2</sup>, *Michael Zabarankin*<sup>2</sup>

## RESEARCH REPORT # 2002-9

Risk Management and Financial Engineering Lab  
Center for Applied Optimization  
Department of Industrial and Systems Engineering  
University of Florida, Gainesville, FL 32611

**Version: December 21, 2002**

*Correspondence should be addressed to: Stanislav Uryasev*

### Abstract

Linear regression is traditionally based on the minimization of variance, or equivalently, standard deviation, but other approaches are possible in which standard deviation is replaced by something more general. A one-to-one correspondence is now known between risk measures, such as have been introduced for various applications in finance, and a large class of deviation measures characterized by simple axioms. Included in that class are asymmetric measures coming from conditional value-at-risk and other currently attractive notions. This paper looks at deviation in that wide sense, formulating the associated problem of regression and investigating the existence and uniqueness of the coefficients that constitute a solution. Such coefficients are characterized in ways that provide a key to their computation.

**Keywords:** *deviation measures, generalized linear regression, coherent risk measures, value-at-risk, conditional value-at-risk, convex analysis.*

---

<sup>1</sup>University of Washington, Department of Mathematics, Box 354350, Seattle, WA 98195-4350;  
E-mail: [rtr@math.washington.edu](mailto:rtr@math.washington.edu). Research supported by NSF grant *DMS-0104055*

<sup>2</sup>University of Florida, ISE Department, P.O. Box 116595, 303 Weil Hall, Gainesville, FL 32611-6595;  
E-mails: [uryasev@ufl.edu](mailto:uryasev@ufl.edu), [zabarank@ufl.edu](mailto:zabarank@ufl.edu); URLs: [www.ise.ufl.edu/uryasev](http://www.ise.ufl.edu/uryasev), <http://plaza.ufl.edu/zabarank>

# 1 Introduction

In linear regression at its most elementary level, a random variable  $Y$  is approximated in terms of another random variable  $X$  by an expression  $cX + d$ , with the values of  $c$  and  $d$  being chosen to minimize the expectation of  $(Y - [cX + d])^2$ . That leads to the familiar picture of the “best line through a swarm of points” when a finite number of  $(X, Y)$  values have been generated as empirical data.

More broadly, one can contemplate approximating  $Y$  in terms of several better-understood random variables  $X_1, \dots, X_n$  by an expression  $c_1X_1 + \dots + c_nX_n + d$ . This is still linear regression, because only a linear combination is involved in the approximation. Although it is customary to think of determining the desired coefficients in that case by solving a system of linear equations in variances and covariances, the framework may also be taken to be one of optimization. Because the expectation of  $(Y - [c_1X_1 + \dots + c_nX_n + d])^2$ , to be minimized, can be written as

$$\sigma^2(Y - [c_1X_1 + \dots + c_nX_n + d]) + (E[Y - [c_1X_1 + \dots + c_nX_n + d]])^2,$$

where the variance term does not actually depend on  $d$ , the regression coefficients can equally well be viewed as solving the problem

$$(\mathcal{P}_0) \quad \text{minimize } \sigma(Y - [c_1X_1 + \dots + c_nX_n + d]) \text{ subject to } E[c_1X_1 + \dots + c_nX_n + d] = EY,$$

or indeed as obtained by minimizing  $\sigma([Y - EY] - c_1[X_1 - EX_1] - \dots - c_n[X_n - EX_n])$  to get  $c_1, \dots, c_n$  and then taking  $d = EY - c_1EX_1 - \dots - c_nEX_n$ .

Minimizing  $\sigma$  is the same of course as minimizing  $\sigma^2$ , but our aim in this paper is to explore the replacement of *standard* deviation  $\sigma$  in  $(\mathcal{P}_0)$  by other measures of the overall extent that a random variable deviates from its expectation. Such *general* deviation measures, having some of the main features of standard deviation such as convexity and linearity to scale, but differing significantly in other ways, may offer important advantages in some applications.

For all of its classical appeal and tractability, standard deviation suffers from serious drawbacks in a number of major applications, particularly in finance. It fails to reflect the “fat tails” of the losses that seem so typical in portfolio performance and insurance, and mean loss variables cannot safely be viewed as normally distributed. Furthermore, it is indifferent to the distinction between “ups and downs,” whereas companies and their clients are usually much more sensitive to “downs.”

Alternatives to standard deviation are readily available now to counter such troubles, and the roster has recently been extended substantially. In financial engineering, nonstandard deviation measures constructed from “conditional value-at-risk” have emerged as especially attractive for their powerful mathematical properties, related to coherency in risk assessment and accessibility in computation. In our paper [8], we have furnished an axiomatic approach to deviation measures including these, and many others. Such measures can even take account of more than just the distribution of a random variable  $X$  by tailoring penalties to the states in which undesirable outcomes occur, as well as to the sizes of the accompanying shortfalls.

How far can a generalization of linear regression beyond the framework of standard deviation be carried, and what new tools might it provide for dealing with random variables that may be far from normally distributed? This is the question we want to start answering here. Already in [8], we have dealt with the fundamentals of optimization problems involving deviation measures. We have laid out the motivation for such problems and characterized their solutions through the methodology of convex analysis. The goal now is to apply those results specifically to linear regression, while enlarging their scope, as needed.

How would one select a particular nonstandard deviation measure to be used in a given application? That is a broad issue in which we can only, at this stage, see glimmerings of insights. It is an issue already familiar in statistics in considerations of the merits of mean absolute deviation over standard deviation, for instance. The theory in [8] ties deviation measures, in general, to *risk measures*, which express attitudes toward the future that can mediate between worst-case analysis and simple reliance on expectations through the designation of a “risk envelope.” The choice of a deviation measure must, in that sense, be derivable from user preferences as an expression of such attitudes. It can be compared that way to the choice of a utility function, although it would be premature to draw strong connections between those two kinds of choices. In any case we can hope that, by working on the details of what happens when standard deviation is replaced by nonstandard deviation in linear regression, we might be able to contribute to a better understanding of how choices should be made.

Many other questions would also need to be answered from the statistical side, before regression with generalized deviation measures could be regarded as more than a proposal undergoing analysis. What can be said about statistical estimates for deviations or relative deviations, and the distributions of such estimates? And what about consistency in sampling, and the development of tests for validating the approximations generated by regression calculations? Ideas abound for how one might be able to build theory in such directions, but of necessity this paper is limited to laying only a foundation.

## 2 Generalized Deviation

Throughout this endeavor, we keep to the framework in which a random variable (r.v.)  $X$  is regarded as a function on a space  $\Omega$  of future states  $\omega$ . We suppose that a probability measure  $P$  for those states is at hand and thereby take  $\Omega$  to be a probability space (supplied in technical terms with a field of “measurable sets” on which  $P$  is defined). We focus on the corresponding  $\mathcal{L}^2$  space consisting of all (“measurable”)  $X$  for which  $E[X^2] = \int_{\Omega}[X(\omega)]^2 dP(\omega)$  is finite. We write  $X \geq X'$  when, in outcome,  $X(\omega) \geq X'(\omega)$  holds almost surely, i.e., is violated, if at all, only with probability 0. We use  $C$  equally to denote a number in  $\mathbb{R} = (-\infty, \infty)$  or the corresponding constant r.v. in  $\mathcal{L}^2$ . The *essential* infimum and supremum of  $X$  will be denoted simply by  $\inf X$  and  $\sup X$ :

$$\inf X = \begin{cases} \text{highest } C \text{ such that } C \leq X & \text{if a constant } C \leq X \text{ exists,} \\ -\infty & \text{otherwise,} \end{cases}$$

$$\sup X = \begin{cases} \text{lowest } C \text{ such that } C \geq X & \text{if a constant } C \geq X \text{ exists,} \\ \infty & \text{otherwise.} \end{cases}$$

In this context, the notion of deviation that we introduced in [8] can be recalled.

**Definition 1** (general deviation measures). *By a deviation measure on  $\mathcal{L}^2$  will be meant any functional  $\mathcal{D} : \mathcal{L}^2 \rightarrow [0, \infty]$  satisfying*

- (D1)  $\mathcal{D}(X + C) = \mathcal{D}(X)$  for all  $X$  and constants  $C$ ,
- (D2)  $\mathcal{D}(0) = 0$ , and  $\mathcal{D}(\lambda X) = \lambda \mathcal{D}(X)$  for all  $X$  and all  $\lambda > 0$ ,
- (D3)  $\mathcal{D}(X + X') \leq \mathcal{D}(X) + \mathcal{D}(X')$  for all  $X$  and  $X'$ ,
- (D4)  $\mathcal{D}(X) > 0$  for all nonconstant  $X$ , whereas  $\mathcal{D}(X) = 0$  for constant  $X$ .

*A deviation measure  $\mathcal{D}$  is said to be coherent if it further satisfies*

- (D5)  $\mathcal{D}(X) \leq EX - \inf X$  for all  $X$ .

In the first place, these axioms ensure through D1 and D4 that  $\mathcal{D}(X)$  depends only on  $X - EX$ , the “uncertain” part of  $X$  expressing the extent to which  $X$  is nonconstant, and that  $\mathcal{D}(X)$  puts

a penalty on such uncertainty. (The assertion for constant r.v.'s in D4 is redundant, being already covered by D1, but has been included for clarity.)

The properties in D2 and D3 are individually known as the *positive homogeneity* and *subadditivity* of a functional on  $\mathcal{L}^2$ ; positive homogeneity is “linearity to scale.” They combine as the property of *sublinearity*, which implies convexity — a crucial support for problems of optimization. Along with D1 and D4, axioms D2 and D3 say that  $\mathcal{D}$  acts much like a “norm” on the space

$$\mathcal{L}_0^2 = \{X \in \mathcal{L}^2 \mid EX = 0\}. \quad (1)$$

The key difference with that classical concept in the mathematics of approximation, is that  $\mathcal{D}$  need not be symmetric:  $\mathcal{D}(-X)$  need not agree with  $\mathcal{D}(X)$ . That feature, of course, is demanded by any approach in which “lower errors” might be rated worse than “upper errors.”

Axiom D5 is a by-product of the general theory of risk measures, pioneered by Artzner, Delbaen, Eber and Heath [2]. They argued that risk measures also had to have a certain property of monotonicity in order to behave in a manner that could be deemed coherent. In [8], we showed that deviation measures  $\mathcal{D}$  satisfying D1, D2, D3 and D4 correspond one-to-one with (what we called) expectation-bounded risk measures  $\mathcal{R}$  on  $\mathcal{L}^2$ . Furthermore, we demonstrated that the coherency of  $\mathcal{R}$  in the sense of [2] thereby translated to  $\mathcal{D}$  satisfying D5. Note that  $EX - \inf X$  measures the “lower range” of  $X$ .

Serving as first examples of deviation measures, obviously, are the standard deviation  $\mathcal{D}(X) = \sigma(X)$  and semideviations  $\mathcal{D}(X) = \sigma_+(X)$  and  $\mathcal{D}(X) = \sigma_-(X)$ , where

$$\sigma_+(X) = (E[\max\{0, X - EX\}^2])^{1/2}, \quad \sigma_-(X) = (E[\max\{0, EX - X\}^2])^{1/2}. \quad (2)$$

Of these three, only  $\sigma_-$  is coherent. Other examples of deviation measures in a similar vein of simple penalties can be obtained from the formula

$$\mathcal{D}(X) = (E[(a \max\{0, X - EX\} + b \max\{0, EX - X\})^p])^{1/p} \quad (3)$$

with  $p \in [1, \infty)$ ,  $a \geq 0$ ,  $b \geq 0$ ,  $a + b > 0$ .

We established in [8, Example 6] that this formula yields a deviation measure as long as  $a \geq 0$ ,  $b \geq 0$ , and  $a + b > 0$ , which moreover is coherent when  $a = 0$  and  $b \leq 1$ , or when  $p = 1$  and merely  $a + b \leq 1$ , but not coherent otherwise (in state spaces  $\Omega$  containing subsets of arbitrarily small positive probability).

Our project of generalized linear regression is not really propelled by such examples of deviation measures as much as it is by others of a rather different sort, which appear to have strong potential in finance and elsewhere. To describe those examples, we need to recall the notion of the *lower  $\alpha$ -tail* of the random variable  $X - EX$  at a level  $\alpha \in (0, 1)$ . This is rigorously defined in the following manner:

$$\left\{ \begin{array}{l} \text{in terms of the distribution function } \Psi \text{ for } X - EX, \text{ the } \textit{lower } \alpha\text{-tail of } X - EX \\ \text{is the r.v. for which the distribution function is } \Psi_\alpha = \alpha^{-1} \min\{\alpha, \Psi\}. \end{array} \right.$$

(The distribution function  $\Psi_\alpha$  is obtained, in other words, by truncating  $\Psi$  from above at the level of  $\alpha$  and then rescaling to have once more a function that ranges from 0 to 1. This precise definition, developed in [7], accounts properly for the possible presence of a jump in  $\Psi$  at the  $\alpha$ -quantile point.) The  *$\alpha$ -conditional value-at-risk* deviation measure is defined then by

$$\mathcal{D}(X) = \text{CVaR}_\alpha^\Delta(X) = -[\text{expectation of the lower } \alpha\text{-tail of } X - EX]. \quad (4)$$

(The “ $-$ ” converts negative gains into positive losses.) This quantity is identical to the expectation of the *downside* r.v.  $EX - X$  conditional on its outcome lying in the portion of its range at or above

the  $1 - \alpha$  quantile, as long as there is no probability atom at that quantile. For more details, and elucidation of the case where an atom is indeed present, we refer to [7].

Conditional value-at-risk grew out of the shortcomings of another percentile-based deviation-like measure, known as value-at-risk and defined in terms of the distribution function  $\Psi$  for  $X - EX$  by

$$\text{VaR}_\alpha^\Delta(X) = -\inf\{z \mid \Psi(z) > \alpha\}. \quad (5)$$

This lacks coherency and, in particular, convexity (so it fails the tests of Definition 1). In contrast, the measure  $\mathcal{D}(X) = \text{CVaR}_\alpha^\Delta(X)$  satisfies all the axioms D1, D2, D3, D4 and D5. The same holds, moreover, for *mixed* CVaR deviation measures which combine different  $\alpha$ 's. They take the form

$$\mathcal{D}(X) = \lambda_1 \text{CVaR}_{\alpha_1}^\Delta(X) + \cdots + \lambda_m \text{CVaR}_{\alpha_m}^\Delta(X) \quad (6)$$

with weights  $\lambda_i \geq 0$  summing to 1, or more generally

$$\mathcal{D}(X) = \int_0^1 \text{CVaR}_\alpha^\Delta(X) d\lambda(\alpha) \quad (7)$$

for a weighting measure  $\lambda$  (nonnegative, with total weight 1); see [8, Example 5]. For additional insights into this class of deviation measures and their ‘‘spectral’’ modes of expression, we refer also to [1].

It should be observed that Definition 1 allows a deviation measure  $\mathcal{D}$  to have  $\mathcal{D}(X) = \infty$  for some r.v.'s  $X$ . Infinite values, if present, are to be handled in D2 and D3 by the conventions that  $\alpha + \infty = \infty$  for any  $\alpha \in (-\infty, \infty]$ , and  $\lambda\infty = \infty$  for any  $\lambda > 0$ , whereas  $0\infty = 0$ . The CVaR-type deviation measures in (4), (6), (7), are finite on  $\mathcal{L}^2$  and thus avoid this eventuality, but infinite values can occur in the preceding examples with exponent  $p$  when  $p > 2$  and there are infinitely many future states in the space  $\Omega$ . Other notable examples of deviation measures that can take on  $\infty$  are

$$\mathcal{D}(X) = EX - \inf X, \quad \text{or} \quad \mathcal{D}(X) = \sup X - EX, \quad \text{or} \quad \mathcal{D}(X) = \sup X - \inf X. \quad (8)$$

They too satisfy D1, D2, D3 and D4, but only the first of them also satisfies D5 (furnishing coherency), and only the third is symmetric. The first measures the lower range of  $X$ , the second measures the upper range, and the third measures the entire range.

A crucial role in understanding general deviation measures, and later in characterizing the coefficients in generalized linear regression, is played by the following concept.

**Definition 2** (risk envelopes). *A subset  $\mathcal{Q}$  of  $\mathcal{L}$  will be called a risk envelope if it has the properties that*

- (Q1)  $\mathcal{Q}$  is a closed, convex set containing 1 (constant r.v.),
- (Q2) every  $Q \in \mathcal{Q}$  has  $EQ = 1$ ,
- (Q3) there is no nonconstant  $X \in \mathcal{L}^2$  such that  $E[Q(EX - X)] \leq 0$  for all  $Q \in \mathcal{Q}$ .

*It will be called a coherent risk envelope if it satisfies, in addition,*

- (Q4)  $Q \geq 0$  for all  $Q \in \mathcal{Q}$ .

In the coherent case where Q4 holds, we can interpret  $\mathcal{Q}$  as a set of *densities*  $Q$  with respect to  $P$  of probability measures  $QdP$  regarded as alternatives to  $P$  (which corresponds to  $Q = 1$ ). Then  $E[Q(EX - X)]$  is the expectation  $E_Q[EX - X] = \int_\Omega [EX - X(\omega)]Q(\omega)dP(\omega)$  with respect to that probability measure. In that context, Q3 insists that, for each nonconstant  $X$ , there must be at least one  $Q \in \mathcal{Q}$  such that  $E_Q X < EX$ . In general, regardless of coherency, we always have

$$E[Q(EX - X)] = \text{covar}(Q, -X) \quad (9)$$

and can think of this covariance as relating the “shape” of  $Q$  to the “the losses under  $X$ .”

The closedness in Q1 refers of course to the usual notion of closedness for subsets in  $\mathcal{L}^2$ : a subset is closed if and only if it contains the limit of any convergent sequence that lies with in it, and for  $\mathcal{L}^2$  the convergence of a sequence  $X_1, X_2, \dots$ , to  $X$  means that  $E[(X_k - X)^2] \rightarrow 0$  as  $k \rightarrow \infty$ .

The explanation of how risk envelopes relate to deviation measures hinges on a semicontinuity property. As a functional on  $\mathcal{L}^2$ , a deviation measure  $\mathcal{D}$  is *lower semicontinuous* if its lower level sets, i.e., the sets having the form  $\{X \mid \mathcal{D}(X) \leq r\}$  for some number  $r$ , are closed in  $\mathcal{L}^2$ . It is *upper semicontinuous* if it satisfies the parallel condition with  $\leq r$  replaced by  $\geq r$ . Continuity is the combination of lower semicontinuity with upper semicontinuity. All of the examples of deviation measures  $\mathcal{D}$  that have been mentioned above are lower semicontinuous, and moreover when those measures do not take on the value  $\infty$  anywhere on  $\mathcal{L}^2$ , they are continuous; see [8, Propositions 1,2].

**Theorem 1** (risk envelope characterization of deviation measures; cf. [8, Theorem 3]). *Risk envelopes  $\mathcal{Q}$  correspond one-to-one with the deviation measures  $\mathcal{D}$  that are lower semicontinuous. Specifically, such a deviation measure has a unique representation of the form*

$$\mathcal{D}(X) = \sup_{Q \in \mathcal{Q}} E[Q(EX - X)] = \sup_{Q \in \mathcal{Q}} \text{covar}(Q, -X) \quad (10)$$

with respect to some risk envelope  $\mathcal{Q}$ , which moreover can be recaptured from  $\mathcal{D}$  by

$$\begin{aligned} \mathcal{Q} &= \{Q \mid E[(1 - Q)X] \leq \mathcal{D}(X) \text{ for all } X\} \\ &= \{Q \mid E[Q(EX - X)] \leq \mathcal{D}(X) \text{ for all } X, EQ = 1\}. \end{aligned} \quad (11)$$

In this correspondence,  $\mathcal{D}$  is coherent if and only if  $\mathcal{Q}$  is coherent.

In the coherent case, the representation of  $\mathcal{D}$  in (10) says that  $\mathcal{D}(X)$  gives the worst-case expectation of the downside variable  $EX - X$  with respect to the probability distributions represented by the densities  $Q \in \mathcal{Q}$ :

$$\mathcal{D}(X) = \sup_{Q \in \mathcal{Q}} E_Q[EX - X]. \quad (12)$$

This interpretation is the source of the “risk envelope” terminology for  $\mathcal{Q}$ , introduced in [8].

The risk envelopes  $\mathcal{Q}$  corresponding to various choices of deviation measures  $\mathcal{D}$ , such as above, were identified in [8]. We list here three prominent cases.

**Example 1** (risk envelopes for particular deviation measures).

- (a)  $\mathcal{D}(X) = \sigma(X)$  corresponds to  $\mathcal{Q} = \{Q \mid \sigma(Q) \leq 1, EQ = 1\}$ .
- (b)  $\mathcal{D}(X) = EX - \inf X$  corresponds to  $\mathcal{Q} = \{Q \mid Q \geq 0, EQ = 1\}$ .
- (c)  $\mathcal{D}(X) = \text{CVaR}_\alpha^\Delta(X)$  corresponds to  $\mathcal{Q} = \{Q \mid 0 \leq Q \leq \alpha^{-1}, EQ = 1\}$ .

The lack of nonnegativity of  $Q$  in Example 1(a) reflects the fact that standard deviation, in contrast to the other two, is not a *coherent* deviation measure.

Under the correspondence in Theorem 1, deviation measures  $\mathcal{D}$  that are finite on  $\mathcal{L}^2$  are associated with risk envelopes  $\mathcal{Q}$  that are bounded in  $\mathcal{L}^2$ , and hence “weakly compact.” (This is true because, for a closed *convex* subset of  $\mathcal{L}^2$ , here  $\mathcal{Q}$  through Q1, boundedness is equivalent not only to weak compactness, but also to linear boundedness — the property that any linear functional has finite infimum and supremum over  $\mathcal{Q}$ ; and that is equivalent to having  $\mathcal{D}(X) < \infty$  for every  $X$  in (10)).

**Example 2** (risk envelopes for deviation mixtures). *Suppose  $\mathcal{D}_1, \dots, \mathcal{D}_m$  are deviation measures that are lower semicontinuous and have  $\mathcal{Q}_1, \dots, \mathcal{Q}_m$  as their associated risk envelopes. Let*

$$\mathcal{D} = \lambda_1 \mathcal{D}_1 + \dots + \lambda_m \mathcal{D}_m \text{ for weights } \lambda_i \geq 0 \text{ with } \lambda_1 + \dots + \lambda_m = 1. \quad (13)$$

Then  $\mathcal{D}$ , which is again a deviation measure that is lower semicontinuous, has its risk envelope  $\mathcal{Q}$  given by

$$\mathcal{Q} = \{ \lambda_1 Q_1 + \dots + \lambda_m Q_m \mid Q_1 \in \mathcal{Q}_1, \dots, Q_m \in \mathcal{Q}_m \}, \quad (14)$$

provided for instance that the measures  $\mathcal{D}_i$  are finite on  $\mathcal{L}^2$ , except for perhaps one of them.

**Detail.** The fact that  $\mathcal{D}$  is another lower semicontinuous deviation measure was recorded already in [8], but the formula for associated risk envelope  $\mathcal{Q}$  was not developed there. It is easy to see that the set  $\mathcal{Q}$  given by this formula does yield  $\mathcal{D}$  through (10), and that  $\mathcal{Q}$  inherits Q1, Q2 and Q3 of Definition 2, except perhaps for the closedness in Q1. It is for the sake of that closedness that the assumption about all but perhaps one of the  $\mathcal{D}_i$ 's being finite comes in. That assumption guarantees that all but perhaps one of the convex sets  $\mathcal{Q}_i$  are weakly compact. It is known in convex analysis that a weighted combination of closed convex sets  $\mathcal{Q}_i$  as in the formula for  $\mathcal{Q}$ , with all but one of them weakly compact, yields a closed set. Therefore, we do have  $\mathcal{Q}$  closed. Since the correspondence in Theorem 1 is one-to-one, it follows that  $\mathcal{Q}$  must be the risk envelope associated with  $\mathcal{D}$ .  $\square$

Deviation measures in the sense of Definition 1, as functionals on  $\mathcal{L}^2$ , encompass more modeling possibilities than might at first be apparent. They can bring into the picture other considerations about  $X$  than just the distribution of  $X$ .

**Definition 3** (distribution-based measures). *A deviation measure  $\mathcal{D}$  will be called distribution-based if  $\mathcal{D}(X)$  depends only on the distribution of  $X$ , or in other words, if  $\mathcal{D}(X_1) = \mathcal{D}(X_2)$  whenever  $X_1$  and  $X_2$  are elements of  $\mathcal{L}^2$  yielding the same distribution function on  $(-\infty, \infty)$ .*

All the specific deviation measures that have been contemplated so far obey this principle.

**Proposition 1** (identification of distribution-based classes).

(a) *Standard deviation and the semideviation measures in (2) are distribution-based, as are the corresponding absolute deviation versions and the other such generalizations in (3).*

(b) *The range-based deviation measures in (8) are all distribution-based.*

(c) *All the CVaR and mixed-CVaR deviation measures in (4), (6) and (7) are distribution-based.*

**Proof.** In (a), confirmation comes from observing that  $\mathcal{D}(X)$  can in every case be written in terms of an integral over  $(-\infty, \infty)$  with respect to the probability measure encoded by the distribution function of  $X$ . In (b),  $\mathcal{D}(X)$  depends only on the support of that probability measure and likewise therefore fits the picture. In (c), it merely needs to be recalled that all CVaR-type deviations are directly defined in terms of distribution functions.  $\square$

From this perspective, it might be imagined that maybe all deviation measures are distribution-based, but that is not a correct conjecture.

**Example 3** (generalized mean absolute deviation, not distribution-based). *Suppose*

$$\mathcal{D}(X) = \int_{\Omega} a(\omega) |X(\omega) - EX| dP(\omega) \text{ for a nonconstant coefficient function } a(\omega) > 0.$$

*Then  $\mathcal{D}$  is a deviation measure that can have  $\mathcal{D}(X_1) \neq \mathcal{D}(X_2)$  when  $X_1$  and  $X_2$  yield the same distribution function on  $(-\infty, \infty)$ , yet differ as elements of  $\mathcal{L}^2$ . Its risk envelope is*

$$\mathcal{Q} = \{ Q \mid |1 - Q(\omega)| \leq a(\omega), EQ = 1 \}.$$

**Detail.** The crucial thing here is that  $\mathcal{D}(X)$  cannot, in general, be expressed by an integral over  $(-\infty, \infty)$ , because it depends on the outcomes of  $a(\omega) |X(\omega) - EX|$  rather than merely the outcomes

of  $X(\omega)$ . To see this in sharp relief, consider the case where  $\Omega$  is the interval  $[0, 6]$  supplied with the uniform probability distribution, and let  $a(\omega) = b$  on  $[0, 3]$  but  $a(\omega) = c$  on  $(3, 6]$ , with  $b > c > 0$ . Let

$$X_1(\omega) = \begin{cases} 2 & \text{on } [0, 2], \\ -1 & \text{on } (2, 6], \end{cases} \quad X_2(\omega) = \begin{cases} -1 & \text{on } [0, 4], \\ 2 & \text{on } (4, 6], \end{cases}$$

observing that both of these r.v.'s take the value 2 with probability  $1/3$  and the value  $-1$  with probability  $2/3$  and thus have the same distribution, moreover with  $EX_1 = EX_2 = 0$ . We calculate that  $\mathcal{D}(X_1) = (1/3)2b + (1/6)b + (1/2)c$ , and on the other hand  $\mathcal{D}(X_2) = (1/2)b + (1/6)c + (1/3)2c$ . This gives  $\mathcal{D}(X_1) - \mathcal{D}(X_2) = (1/3)(b - c) > 0$ , so that  $\mathcal{D}(X_1) \neq \mathcal{D}(X_2)$ .

It is easy to verify that the designated set  $\mathcal{Q}$  satisfies the axioms Q1, Q2 and Q3, and returns  $\mathcal{D}$  through (10). It must therefore, by Theorem 1, be the risk envelope for this measure  $\mathcal{D}$ .  $\square$

Example 3 offers only one out of a host of possibilities. Nonetheless, it clearly brings home the point that deviation measures may depart from the distribution-based category because of taking the ‘‘scenarios’’  $\omega \in \Omega$  themselves into account, and not just the outcomes  $X(\omega)$ . This is a previously unexploited notion with interesting potential for applications, because it allows refinements in the modeling of user-preferences. An individual’s reaction to a shortfall  $EX - X(\omega) > 0$  might, for instance, depend not only on the size of this shortfall but also on the kind of scenario in which it occurs. Some scenario ranges could be more painful or distressing than others, as when one’s own interests suffer while those of one’s competitors are prospering.

### 3 Regression Model

Equipped with this wealth of deviation measures alongside of the standard ones, we wish to explore correspondingly different versions of linear regression. To be able to utilize the relationships with risk envelopes in Theorem 1, we assume throughout the rest of this paper that

*$\mathcal{D}$  is a deviation measure which is lower semicontinuous.*

Coherency is not demanded (we assume only D1, D2, D3 and D4, not necessarily D5), because that would exclude standard deviation, which must be maintained as an essential example. Coherent measures such as the CVaR-type deviations are among the primary targets for new applications, however.

We fix a collection of reference r.v.’s  $X_1, \dots, X_n$  in  $\mathcal{L}^2$  and, for arbitrary  $Y \in \mathcal{L}^2$ , try to approximate  $Y$  by a combination  $c_1X_1 + \dots + c_nX_n + d$  in a manner analogous to classical linear regression, as crystalized from the optimization standpoint in problem  $(\mathcal{P}_0)$  of the introduction. The key is the replacement of the standard deviation  $\sigma$  in the optimization problem  $(\mathcal{P}_0)$ , corresponding to standard linear regression, by the deviation measure  $\mathcal{D}$ .

**Definition 4** (generalized regression coefficients). *A choice of  $c_1, \dots, c_n, d$ , will be said to constitute regression coefficients for  $Y$  with respect to  $X_1, \dots, X_n$  and  $\mathcal{D}$  if it solves the optimization problem*

$$(\mathcal{P}) \quad \text{minimize } \mathcal{D}(Y - [c_1X_1 + \dots + c_nX_n + d]) \text{ subject to } E[c_1X_1 + \dots + c_nX_n + d] = EY,$$

which is equivalent to saying that

- (a)  $(c_1, \dots, c_n)$  minimizes  $\mathcal{D}([Y - EY] - c_1[X_1 - EX_1] - \dots - c_n[X_n - EX_n])$ ,
- (b)  $d = EY - c_1EX_1 - \dots - c_nEX_n$ .

When  $c_1, \dots, c_n$  and  $d$  are regression coefficients, the random variable  $X = c_1X_1 + \dots + c_nX_n + d$  can be regarded as a *closest approximation* to  $Y$  from the linear subspace  $\mathcal{X}$  of  $\mathcal{L}^2$  generated by  $X_1, \dots, X_n$  and the constant r.v.'s (subject to the equal expectation constraint). It is a  $\mathcal{D}$ -*projection* of  $Y$  on  $\mathcal{X}$ . We have to say “a” closest approximation, and “a” projection, because there is no guarantee in general that the coefficients are uniquely determined. Uniqueness could anyway often be the case for particular  $X_1, \dots, X_n$ , or for special choices of  $\mathcal{D}$ . The uniqueness issue will further be addressed below, along with existence.

Of course, because  $\mathcal{D}$  might not be symmetric, the minimization of  $\mathcal{D}(Y - [c_1X_1 + \dots + c_nX_n + d])$  in  $(\mathcal{P})$  might not lead to the same answer as the minimization of  $\mathcal{D}([c_1X_1 + \dots + c_nX_n + d] - Y)$ . The latter corresponds, however, to minimizing  $\tilde{\mathcal{D}}(Y - [c_1X_1 + \dots + c_nX_n + d])$  for the functional  $\tilde{\mathcal{D}}(X) = \mathcal{D}(-X)$ . Since  $\tilde{\mathcal{D}}$  is another deviation measure, likewise satisfying D1, D2, D3 and D4, and inheriting lower semicontinuity from  $\mathcal{D}$ , both minimizations will be covered by the results to be presented.

Nonetheless, the modeling decision about which expression ought to be minimized could be important in some situations. The asymmetric deviation measures that are coherent scheme give priority to downside penalization. Therefore, in focusing on  $\mathcal{D}(Y - X)$  instead of  $\mathcal{D}(X - Y)$ , in our framework where  $X = c_1X_1 + \dots + c_nX_n + d$  and  $EX = EY$ , we are adopting the position that *overestimates*, corresponding to outcomes with  $X(\omega) > Y(\omega)$  (i.e., negative outcomes of  $Y - X$ ), are potentially more worrisome than *underestimates*, corresponding to outcomes with  $X(\omega) < Y(\omega)$  (i.e., positive outcomes of  $Y - X$ ).

That orientation is appropriate in finance when  $X$  and  $Y$  represent uncertain returns, and shortfall is the dominating concern. The fact that nonstandard deviation measures have the capability making such a distinction is one of their major attractions.

**Proposition 2** (properties of the regression problem). *The nonnegative function*

$$\begin{aligned} f(c_1, \dots, c_n) &= \mathcal{D}(Y - [c_1X_1 + \dots + c_nX_n + d]) \\ &= \mathcal{D}([Y - EX] - c_1[X_1 - EX_1] - \dots - c_n[X_n - EX_n]) \end{aligned}$$

being minimized over  $\mathbb{R}^n$  in problem  $(\mathcal{P})$  is convex. Furthermore, it is lower semicontinuous, in fact continuous when it does not take on  $\infty$  anywhere, as is true in particular when  $\mathcal{D}$  only has finite values on  $\mathcal{L}^2$ . Thus, any local minimum in  $\mathcal{P}$  is a global minimum, and the minimizing elements (if any) form a closed, convex set.

**Proof.** The convexity and lower semicontinuity of  $f$  follow immediately from the convexity and lower semicontinuity of  $\mathcal{D}$  as a functional on  $\mathcal{L}^2$ . Any finite convex function on  $\mathbb{R}^n$  is continuous; cf. [4, Cor. 10.1.1]. The remaining assertions draw on the fact that, when  $f$  is convex and lower semicontinuous, all sets of the form  $\{(c_1, \dots, c_n) \mid f(c_1, \dots, c_n) \leq r\}$  are convex and closed; when  $r$  is the minimum value in the problem, this yields the set of optimal solutions. The fact that a local minimum of a convex function is a global minimum is one of the best-known features of convex optimization.  $\square$

The closedness and convexity of the solution set to  $\mathcal{P}$  implies that the corresponding  $\mathcal{D}$ -projections  $c_1X_1 + \dots + c_nX_n + d$  of  $Y$  on  $\mathcal{X}$  form a closed, convex subset of the subspace  $\mathcal{X}$  in  $\mathcal{L}^2$ . Still unsettled, however, are the questions of whether these sets are nonempty, and whether they can contain more than one element. These questions need to be addressed separately.

**Theorem 2** (existence of regression coefficients). *As long as there is at least one choice of  $X = c_1X_1 + \dots + c_nX_n + d$  such that  $\mathcal{D}(Y - X) < \infty$ , problem  $(\mathcal{P})$  will have a solution, i.e., regression coefficients for  $Y$  with respect to  $X_1, \dots, X_n$  and  $\mathcal{D}$  will exist.*

**Proof.** Through the prescription in Definition 4, the existence of a solution to  $(\mathcal{P})$  corresponds to the function  $f$  in Proposition 2 attaining a minimum on  $\mathbb{R}^n$ . We know, of course, that  $f(c_1, \dots, c_n) \geq 0$  for all  $(c_1, \dots, c_n) \in \mathbb{R}^n$  (by axiom D4), and on the other hand that  $f(c_1, \dots, c_n) < \infty$  for at least one choice of  $(c_1, \dots, c_n) \in \mathbb{R}^n$  (by the assumption that begins the statement of the theorem), so that  $0 \leq \inf f < \infty$ . However, to be sure of the existence of some  $(c_1, \dots, c_n) \in \mathbb{R}^n$  such that  $f(c_1, \dots, c_n) = \inf f$ , we need something more. Because  $f$  is convex and lower semicontinuous, we can invoke a sufficient condition in [4, Corollary 27.3.3], involving so-called directions of recession. The verification of this condition requires demonstrating that if  $(\bar{c}_1, \dots, \bar{c}_n)$  and  $(c'_1, \dots, c'_n)$  are such that

$$f(\bar{c}_1 + \lambda c'_1, \dots, \bar{c}_n + \lambda c'_n) \leq f(\bar{c}_1, \dots, \bar{c}_n) < \infty \text{ for all } \lambda \in (0, \infty), \quad (15)$$

then

$$f(\bar{c}_1 + \lambda c'_1, \dots, \bar{c}_n + \lambda c'_n) = f(\bar{c}_1, \dots, \bar{c}_n) \text{ for all } \lambda \in (-\infty, \infty), \quad (16)$$

Here, (15) translates through the definition of  $f$  and the positive homogeneity property D2 of  $\mathcal{D}$  (and the insensitivity to constant shifts in D1) to having

$$\mathcal{D}(\lambda^{-1}Y - (c'_1 + \lambda^{-1}\bar{c}_1)X_1 - \dots - (c'_n + \lambda^{-1}\bar{c}_n)X_n) \leq \lambda^{-1}\mathcal{D}(Y - \bar{c}_1X_1 - \dots - \bar{c}_nX_n)$$

for all  $\lambda \in (0, \infty)$ . Taking the limit as  $\lambda \rightarrow \infty$  (with everything else fixed) and utilizing the nonnegativity and lower semicontinuity of  $\mathcal{D}$ , we see that it implies  $\mathcal{D}(-c'_1X_1 - \dots - c'_nX_n) = 0$ , and hence through D4 that  $-c'_1X_1 - \dots - c'_nX_n \equiv C$  for some constant  $C$ . Then, by D1,

$$\begin{aligned} f(\bar{c}_1 + \lambda c'_1, \dots, \bar{c}_n + \lambda c'_n) &= \mathcal{D}(Y - \bar{c}_1X_1 - \dots - \bar{c}_nX_n + \lambda C) \\ &= \mathcal{D}(Y - \bar{c}_1X_1 - \dots - \bar{c}_nX_n) = f(\bar{c}_1, \dots, \bar{c}_n), \end{aligned}$$

so we do have (16), as required.  $\square$

What can be said about the uniqueness of a solution? An elementary difficulty could be an interdependence of the reference r.v.'s  $X_j$ . If there are different choices of coefficients,  $c_1, \dots, c_n, d$ , and on the other hand  $c'_1, \dots, c'_n, d'$ , such that the r.v.'s  $X = c_1X_1 + \dots + c_nX_n + d$  and  $X' = c'_1X_1 + \dots + c'_nX_n + d'$  are identical, both choices might give the minimum, and both would then furnish regression coefficients for  $Y$  with respect to  $X_1, \dots, X_n$ . But in this case the  $\mathcal{D}$  projection of  $Y$  on  $\mathcal{X}$  would at least be unique.

Is it possible to have non-uniqueness in  $\mathcal{D}$ -projections as well? To thoroughly eliminate that, a further assumption on  $\mathcal{D}$  can be brought in.

**Definition 5** (strict deviation measures). *A deviation measure  $\mathcal{D}$  will be called strict if it satisfies on top of the subadditivity axiom D3, the stronger condition:*

(D3')  $\mathcal{D}(X + X') < \mathcal{D}(X) + \mathcal{D}(X')$  when  $\mathcal{D}(X) < \infty$  and  $\mathcal{D}(X') < \infty$ , unless  $X$  or  $X'$  is constant, or  $X' = cX + d$  for some  $c > 0$  and  $d$ .

Of course, to say that  $X' = cX + d$  for some constants  $c > 0$  and  $d$  is to say that  $X = c'X' + d'$  for some constants  $c' > 0$  and  $d'$ , with  $c' = 1/c$  and  $d' = -d/c$ , so it is not necessary to include both conditions in D3' for the sake of symmetry. The provision in D3' for an exception when  $X$  or  $X'$  is constant (but perhaps not both) is not covered by this symmetric relationship.

In the cases where D3' allows an exception to strict inequality, we anyway have, merely on the basis of D1 and D2, that  $\mathcal{D}(X + X') = \mathcal{D}(X) + \mathcal{D}(X')$ . In that sense, D3 is implied by D3' under D1 and D2.

**Proposition 3** (strict convexity property). *A deviation measure  $\mathcal{D}$  is strict if and only if  $\mathcal{D}^2$  is strictly convex on the pure uncertainty subspace  $\mathcal{L}_0^2$  of  $\mathcal{L}$  in (1):*

$$\begin{aligned} \mathcal{D}((1-\lambda)X + \lambda X')^2 &< (1-\lambda)\mathcal{D}(X)^2 + \lambda\mathcal{D}(X')^2 \text{ for } 0 < \lambda < 1 \\ &\text{when } X, X' \in \mathcal{L}_0^2, X \neq X', \mathcal{D}(X) < \infty, \mathcal{D}(X') < \infty. \end{aligned} \quad (17)$$

**Proof.** Fix any  $X$  and  $X'$  in  $\mathcal{L}_0^2$  with  $\mathcal{D}(X) < \infty$  and  $\mathcal{D}(X') < \infty$ . First we assume that  $\mathcal{D}$  is strict and argue toward the strict convexity of  $\mathcal{D}^2$ . Let  $\theta(t) = t^2$ , this being a strictly convex, increasing function on  $[0, \infty)$ . Then  $\mathcal{D}(X)^2 = \theta(\mathcal{D}(X))$ . For  $\lambda \in (0, 1)$  we have by D2 and D3' that

$$\mathcal{D}((1-\lambda)X + \lambda X') \leq \mathcal{D}((1-\lambda)X) + \mathcal{D}(\lambda X') = (1-\lambda)\mathcal{D}(X) + \lambda\mathcal{D}(X')$$

with strict inequality holding whenever the strict inequality in D3' holds. Therefore

$$\theta(\mathcal{D}((1-\lambda)X + \lambda X')) \leq \theta((1-\lambda)\mathcal{D}(X) + \lambda\mathcal{D}(X')) \leq (1-\lambda)\theta(\mathcal{D}(X)) + \lambda\theta(\mathcal{D}(X')),$$

where the first inequality is strict whenever the strict inequality in D3' holds, and the second inequality is strict unless  $\mathcal{D}(X) = \mathcal{D}(X')$ . Therefore, we will have the desired strict inequality in (17) if the assumptions on  $X$  and  $X'$  in (17) preclude  $X$  and  $X'$  from fitting one of the exceptional cases in D3' and simultaneously having  $\mathcal{D}(X') = \mathcal{D}(X)$ .

In the exceptional case in D3' where  $X$  or  $X'$  is constant, the constant would have to be 0 (since these r.v.'s belong to  $\mathcal{L}_0^2$ ). But the condition  $\mathcal{D}(X') = \mathcal{D}(X)$  would imply then through D4 that both  $X$  and  $X'$  are 0, in contradiction to  $X \neq X'$ . Hence, this case is impossible. In the other exceptional case in D3', we have  $X' = cX + d$  for nonconstant  $X$  and some  $c > 0$  and  $d$ , so that  $\mathcal{D}(X') = c\mathcal{D}(X) > 0$  and  $EX' = cEX + d$ . Moreover  $EX = EX' = 0$ , because  $X$  and  $X'$  are in  $\mathcal{L}_0^2$ . That, along with the condition  $\mathcal{D}(X') = \mathcal{D}(X)$ , requires  $c = 1$  and  $d = 0$ , hence again that  $X' = X$ , contrary to assumption. This confirms that (17) holds when  $\mathcal{D}$  is strict.

Conversely, suppose that (17) holds. Consider any  $X$  and  $X'$  in  $\mathcal{L}_0^2$  that do not fall into one of the exceptional cases in D3'. Let  $d = \mathcal{D}(X)$  and  $d' = \mathcal{D}(X')$ ; then  $d > 0$  and  $d' > 0$  by D4. We need to verify that  $\mathcal{D}(X + X') < d + d'$ . Here  $X + X'$  likewise belongs to  $\mathcal{L}_0^2$ .

Let  $\bar{X} = X/d$  and  $\bar{X}' = X'/d'$ , so that  $\mathcal{D}(\bar{X}) = 1$  and  $\mathcal{D}(\bar{X}') = 1$ ; then  $\bar{X} \neq \bar{X}'$ , since otherwise  $X' = cX$  for  $c = d'/d$ , a relationship that has been excluded. Observing that

$$X + X' = t[(1-\lambda)\bar{X} + \lambda\bar{X}'] \text{ for } t = d + d', \lambda = d'/(d + d') \in (0, 1),$$

we argue now that

$$\begin{aligned} \mathcal{D}(X + X')^2 &= \mathcal{D}(t[(1-\lambda)\bar{X} + \lambda\bar{X}'])^2 = t^2\mathcal{D}((1-\lambda)\bar{X} + \lambda\bar{X}')^2 \\ &< t^2[(1-\lambda)\mathcal{D}(\bar{X})^2 + \lambda\mathcal{D}(\bar{X}')^2] = t^2, \end{aligned}$$

where the strict inequality holds by (17), and the final equation holds because  $\mathcal{D}(\bar{X}) = 1$  and  $\mathcal{D}(\bar{X}') = 1$ . Thus  $\mathcal{D}(X + X')^2 < t^2$ , so  $\mathcal{D}(X + X') < t = d + d'$ , which is what we needed to prove.  $\square$

**Theorem 3** (uniqueness of regression coefficients). *Suppose that  $\mathcal{D}$  is strict, and that the reference variables  $X_j$  are free from interdependence in the sense that*

$$\text{no choice of } c_1, \dots, c_n, d, \text{ makes } c_1X_1 + \dots + c_nX_n + d \equiv 0, \text{ except } c_1 = \dots = c_n = d = 0.$$

*Then the regression coefficients for  $Y$  with respect to  $X_1, \dots, X_n$  and  $\mathcal{D}$  are unique.*

**Proof.** The regression problem ( $\mathcal{P}$ ) has been seen to be equivalent to minimizing  $\mathcal{D}$ , or equivalently  $\mathcal{D}^2$ , over the subset  $\mathcal{X}_0$  of  $\mathcal{L}_0^2$  that is generated by the variables  $X_1 - EX_1, \dots, X_n - EX_n$ . The strict

convexity of  $\mathcal{D}^2$  guaranteed by Proposition 3 ensures that, in that problem over  $\mathcal{X}_0$ , there can be at most one solution. On the other hand, our assumption on  $X_1, \dots, X_n$  means that  $X_1 - EX_1, \dots, X_n - EX_n$  are linearly independent as elements of  $\mathcal{L}^2$ , so that whenever an element of  $\mathcal{X}_0$  is represented as  $c_1(X_1 - EX_1) + \dots + c_n(X_n - EX_n)$  the coefficients  $c_1, \dots, c_n$  are uniquely determined. The corresponding  $d$  is fixed then by the formula in (b) of Definition 4.  $\square$

**Example 4** (deviations lacking strictness). *Although standard deviation  $\mathcal{D}(X) = \sigma(X)$  is strict (but incoherent), none of the (coherent) deviation measures  $\mathcal{D}(X) = \sigma_-(X)$ ,  $\mathcal{D}(X) = EX - \inf X$ , or  $\mathcal{D}(X) = \text{CVaR}_\alpha^\Delta(X)$  is strict.*

**Detail.** The strictness of standard deviation can be established directly, but it follows also from Proposition 3 through the fact that the functional  $\mathcal{F}(X) = E[X^2]$  is strictly convex. Strictness fails for  $\mathcal{D}(X) = \sigma_-(X)$  because, for instance, two r.v.'s  $X$  and  $X'$  having  $EX = EX' = 0$  and identical downsides but different upsides,

$$X(\omega) = X'(\omega) < 0 \text{ for } \omega \in \Omega' \subset \Omega, \text{ but otherwise } X(\omega) \geq 0, X'(\omega) \geq 0, X(\omega) \neq X'(\omega),$$

will have  $\sigma_-(X) = \sigma_-(X')$  (since  $\sigma_-$  pays attention only to downsides), and on the other hand  $\sigma_-(X + X') = \sigma_-(2X) = 2\sigma_-(X)$  (since the downside of  $X + X'$  is exactly twice the downside of  $X$ ), and consequently  $\sigma_-(X + X') = \sigma_-(X) + \sigma_-(X')$ .

The counterexample is similar for  $\mathcal{D}(X) = EX - \inf X$ : let  $X$  and  $X'$  have  $EX = EX' = 0$  and achieve their minimum over  $\Omega$  on the same subset  $\Omega'$  (of probability intermediate between 0 and 1), but differ elsewhere.

In the case of  $\mathcal{D}(X) = \text{CVaR}_\alpha^\Delta(X)$ , suppose  $\Omega'$  is a subset of  $\Omega$  having probability  $\alpha$ . Consider the subset  $\mathcal{K}$  of  $\mathcal{L}^2$  consisting of all the r.v.'s  $X$  such that  $X(\omega) = 0$  when  $\omega \in \Omega'$  but  $X(\omega) > 0$  when  $\omega \notin \Omega'$ . For any  $X$  in  $\mathcal{K}$ ,  $X - EX$  achieves its minimum on  $\Omega'$  but nowhere else, this minimum value is  $-EX$  and it is taken on with probability  $\alpha$ , so that  $\text{CVaR}_\alpha^\Delta(X) = EX$ . When  $X$  and  $X'$  both belong to  $\mathcal{K}$ , we get  $X + X'$  in  $\mathcal{K}$  as well, so that  $\mathcal{D}(X + X') = E[X + X'] = EX + EX' = \mathcal{D}(X) + \mathcal{D}(X')$ . This equation holds even though  $X$  and  $X'$  can differ outside of  $\Omega'$ .  $\square$

## 4 Characterization of Regression Coefficients

In standard linear regression, the coefficients  $c_1, \dots, c_n$  can be determined by solving a system of linear equations involving a variance-covariance matrix. Such a simple approach cannot be expected in general, but that does not mean it would necessarily be difficult for the coefficients to be calculated. They can be obtained by solving the optimization problem  $(\mathcal{P})$ , and by now there are many highly refined techniques in optimization to employ for such a purpose. In the case of  $\mathcal{D}(X) = \text{CVaR}_\alpha^\Delta(X)$ , for instance, and even for mixed CVaR deviations, the computations can be reduced to linear programming and carried out quickly and efficiently. The key there is a representation formula brought to light in [6] and developed further in [7].

In view of those possibilities, one should not think of having to determine regression coefficients from “closed form” expressions derived as optimality conditions for problem  $(\mathcal{P})$ . Nevertheless, optimality conditions can be useful in shedding light on the nature of the coefficients and their properties. Such conditions may have theoretical significance, as well.

Optimality has already been characterized in [8, Theorem 5] for deviation problems broader than  $(\mathcal{P})$ . We now want to specialize that result to the situation here. Due to the fact that the function  $f$  being minimized in  $(\mathcal{P})$ , and the deviation measure  $\mathcal{D}$  behind it, need not be differentiable in the

usual sense — so that their gradients may not be well defined everywhere — it is essential to rely on a substitute concept of “subgradient” which comes from convex analysis.

**Definition 6** (deviation subgradients). *An element  $G \in \mathcal{L}^2$  is called a subgradient of a deviation measure  $\mathcal{D}$  at  $X$  if*

$$\mathcal{D}(X') \geq \mathcal{D}(X) + E[G(X' - X)] \text{ for all } X' \in \mathcal{L}^2. \quad (18)$$

The set of all such subgradients is denoted by  $\partial\mathcal{D}(X)$ .

When  $\mathcal{D}$  does have a gradient at  $X$ , that gradient is the unique  $G$  for which the inequality holds, and  $\partial\mathcal{D}(X)$  reduces to the singleton set  $\{G\}$ .

**Proposition 4** (characterization of subgradients). *At any  $X$ , the subgradient set  $\partial\mathcal{D}(X)$  is a convex, closed subset of  $\mathcal{L}^2$ ; when  $\mathcal{D}$  is finite everywhere, it is sure moreover to be nonempty and bounded. In terms of the risk envelope  $\mathcal{Q}$  corresponding to  $\mathcal{D}$ , it is given by the formula*

$$\partial\mathcal{D}(X) = \{1 - Q \mid \text{for some } Q \in \mathcal{Q}_X\} \text{ with } \mathcal{Q}_X = \underset{Q \in \mathcal{Q}}{\operatorname{argmax}} E[Q(EX - X)], \quad (19)$$

where “argmax” refers to the set of elements  $Q$  for which the maximum in question is attained.

**Proof.** This formula for  $\partial\mathcal{D}(X)$  was established in [8, Theorem 5]. The convexity and closedness of  $\partial\mathcal{D}(X)$  were noted in that paper as well (these properties being elementary consequences of the definition). For the nonemptiness and boundedness of  $\mathcal{D}(X)$ , it would be enough to have  $\mathcal{D}$  finite on some neighborhood of  $X$ ; this would hold for any lower semicontinuous convex functional on  $\mathcal{L}^2$ , not just  $\mathcal{D}$ , and is a well known fact in convex analysis.  $\square$

Note that the maximization problem for which  $\mathcal{Q}_X$  is the “argmax” in Proposition 4 is the one which, in Theorem 1, produces the value of  $\mathcal{D}(X)$ .

**Theorem 4** (optimality conditions for regression). *Suppose  $\mathcal{D}$  is finite everywhere on  $\mathcal{L}^2$ . In that case,  $c_1, \dots, c_n$  and  $d$  furnish regression coefficients for  $Y$  with respect to  $X_1, \dots, X_n$  and  $\mathcal{D}$  if and only if ( $d$  is tied to  $c_1, \dots, c_n$  by the equation in (b) of Definition 4 and) there exists  $Q \in \mathcal{L}^2$  such that*

- (a)  $1 - Q \in \partial\mathcal{D}(Z)$  for  $Z = Y - (c_1X_1 + \dots + c_nX_n + d)$ ,
- (b)  $\operatorname{covar}(Q, X_j) = 0$  for  $j = 1, \dots, n$ .

**Proof.** Here  $EZ = 0$ ;  $Z$  can also be expressed as  $(Y - EY) - c_1(X_1 - EX_1) - \dots - c_n(X_n - EX_n)$ , corresponding to the minimization problem as recast in (a) of Definition 4. In terms of the variables

$$Z_0 = Y - EY, \quad Z_1 = -(X_1 - EX_1), \quad \dots, \quad Z_n = -(X_n - EX_n),$$

we can think of this problem as having the form:

$$\text{minimize } \mathcal{D}(c_0Z_0 + c_1Z_1 + \dots + c_nZ_n) \text{ subject to } 1c_0 + 0c_1 + \dots + 0c_n = 1. \quad (20)$$

This fits the pattern of optimization in [8, Theorem 5], which more generally would allow a system of finitely many linear constraints (a possible mixture of equations and inequalities on  $c_0, c_1, \dots, c_n$ ). Let  $\mathcal{Z}$  denote the subspace of  $\mathcal{L}^2$  (actually in  $\mathcal{L}_0^2$ ) consisting of all  $Z = c_0Z_0 + c_1Z_1 + \dots + c_nZ_n$  for coefficients satisfying the constraint in (20); in other words,  $\mathcal{Z}$  consists of all r.v.’s of the form  $Z = (Y - EY) - c_1(X_1 - EX_1) - \dots - c_n(X_n - EX_n)$  (with no condition on  $c_1, \dots, c_n$ ). On the basis of [8, Theorem 5] (in which  $\mathcal{D}$  is assumed finite, as here), the minimum is attained at  $Z$  if and only if there exists  $Q$  such that (a) holds and, in addition,  $Q - 1$  belongs to the “normal cone”  $N_{\mathcal{Z}}(Z)$  to  $\mathcal{Z}$  at  $Z$ . There is no need to go into the details of what the “normal cone” means in the general situation covered by [8, Theorem 5], because the case at hand is very simple.

By invoking the normal cone formula in [8, Proposition 3] with respect to the single-constraint system in (20) that gives  $\mathcal{Z}$ , one sees that  $Q-1 \in N_{\mathcal{Z}}(Z)$  if and only if there is a multiplier  $\lambda \in (-\infty, \infty)$  such that  $E[(1-Q)Z_0] = \lambda$  but  $E[(1-Q)Z_j] = 0$  for  $j = 1, \dots, n$ . The first of these equations drops out, due to the arbitrariness of  $\lambda$ . The others can be identified as the covariance equations in (b).  $\square$

**Example 5** (coverage of standard linear regression). *In the case of Theorem 4 where  $\mathcal{D}(X) = \sigma(X)$  and the residual  $Z = Y - [c_1X_1 + \dots + c_nX_n + d]$  does not entirely vanish, condition (a) means that*

$$1 - Q = Z/\sigma(Z). \quad (21)$$

The conditions in (b) come out then as  $\text{covar}(Z, X_j) = 0$ , which can be expanded as

$$c_1 \text{covar}(X_1, X_j) + \dots + c_n \text{covar}(X_n, X_j) = \text{covar}(Y, X_j) \text{ for } j = 1, \dots, n. \quad (22)$$

This system of  $n$  linear equations in the  $n$  coefficients  $c_j$  can be solved for those coefficients (when the matrix of the system, namely the variance-covariance matrix for  $X_1, \dots, X_n$  is nonsingular).

**Detail.** Here  $\mathcal{D}$  is differentiable at  $Z$  and (21) gives the gradient, which can be equated with  $1 - Q$ . The rest is then obvious.  $\square$

**Example 6** (regression with standard semideviation). *In the case of Theorem 4 where  $\mathcal{D}(X) = \sigma_-(X)$  and the residual  $Z = Y - [c_1X_1 + \dots + c_nX_n + d]$  does not entirely vanish, condition (a) means that*

$$1 - Q = Z_-/\sigma_-(Z_-) \text{ with } Z_- = \min\{0, Z\}. \quad (23)$$

The equations in (b) come out then as  $\text{covar}(Z_-, X_j) = 0$ , but cannot be expanded into a system of linear equations in the unknowns  $c_1, \dots, c_n$  in the manner of (22). Instead, they constitute a system of nonlinear equations in these unknowns,

$$\text{covar}([Y - [c_1X_1 + \dots + c_nX_n + d]]_-, X_j) = 0 \text{ for } j = 1, \dots, n. \quad (24)$$

The determination of the unknowns does not actually require solving that nonlinear system numerically, however (because direct techniques of optimization can be applied to the underlying minimization problem instead).

**Detail.** Everything depends on the claim that the subgradient set  $\partial\mathcal{D}(Z)$  reduces to the unique element in (23). This fact about  $\sigma_-$  was not developed in [8] but can be deduced on general principles of convex analysis. First, we look at the functional  $\mathcal{J} = \frac{1}{2}\mathcal{D}^2 = \frac{1}{2}\sigma_-^2$ , which has the form

$$\mathcal{J}(X) = \int_{\Omega} \varphi(X(\omega) - EX) dP(\omega) \text{ for } \varphi(z) = \frac{1}{2} \max\{0, -z\}^2.$$

Here  $\varphi$  is convex and differentiable on  $\mathbb{R}$  with derivative  $\varphi'(z) = -\max\{0, -z\} = \min\{0, z\}$ . The subgradients of  $\mathcal{J}$  can be determined from a general formula for “integral functionals” in [3]. That formula says that  $\partial\mathcal{J}(X)$  consists of the unique element  $G$  defined by  $G(\omega) = \varphi'(X(\omega) - EX)$  (up to the usual almost sure equivalence in  $\mathcal{L}^2$ ). Here we are applying this with  $X$  taken to be the r.v.  $Z = Y - [c_1X_1 + \dots + c_nX_n + d]$ , which has  $EZ = 0$ , and thereby get  $G = Z_-$ . On the other hand, because  $\mathcal{J} = \frac{1}{2}\mathcal{D}^2$ , one has by a general chain rule of convex analysis that  $\partial\mathcal{J}(Z) = \mathcal{D}(Z)\partial\mathcal{D}(Z)$ . We are assuming here that  $Z \neq 0$ , so the semivariance  $\sigma_-(Z) = \mathcal{D}(Z)$  is positive. The conclusion then is that  $\partial\mathcal{D}(Z)$  consists solely of  $Z_-$  divided by  $\sigma_-(Z_-)$ .  $\square$

**Example 7** (regression for worst-case risk). In the case of Theorem 4 where  $\mathcal{D}(X) = EX - \inf X$ , which is a finite deviation measure on  $\mathcal{L}^2$  when the state space  $\Omega$  is a finite set, condition (a) means that

$$EQ = 1 \text{ with } \begin{cases} Q(\omega) = 0 & \text{if } Z(\omega) > \inf Z, \\ Q(\omega) \geq 0 & \text{if } Z(\omega) = \inf Z. \end{cases} \quad (25)$$

Optimality corresponds to the existence of such  $Q$ , representing a probability density concentrated in the set where  $X$  attains its minimum, such that  $\text{covar}(Q, X_j) = 0$  for  $j = 1, \dots, n$ . Again, these constitute nonlinear equations in the unknowns  $c_1, \dots, c_n$  as parameters in the specification of  $Z$  (but there is no requirement to solve such a nonlinear system numerically).

**Detail.** The subgradient characterization in (25) comes this time from [8, Example 19].  $\square$

**Example 8** (CVaR regression). In the case of Theorem 4 where  $\mathcal{D}(X) = \text{CVaR}_\alpha^\Delta(X)$ , which is always finite everywhere on  $\mathcal{L}^2$ , condition (a) means that

$$EQ = 1 \text{ with } Q(\omega) \begin{cases} = \alpha^{-1} & \text{on } \{\omega \mid Z(\omega) < -\text{VaR}_\alpha^\Delta(Z)\}, \\ \in [0, \alpha^{-1}] & \text{on } \{\omega \mid Z(\omega) = -\text{VaR}_\alpha^\Delta(Z)\}, \\ = 0 & \text{on } \{\omega \mid Z(\omega) > -\text{VaR}_\alpha^\Delta(Z)\}, \end{cases} \quad (26)$$

where  $\text{VaR}_\alpha^\Delta(Z)$  is defined by (5). Optimality corresponds to the existence of such a  $Q$  having  $\text{covar}(Q, X_j) = 0$  for  $j = 1, \dots, n$ , and this can be viewed once more as system of nonlinear equations in the unknowns  $c_1, \dots, c_n$  as parameters in the specification of  $Z$  (but which does not have to be solved numerically from that perspective).

Further insight can be gained by considering the case where there is zero probability of  $Z$  taking on the threshold value  $-\text{VaR}_\alpha^\Delta(Z)$ , so that the distribution function for  $Z$  has no jump at this value, and  $\text{CVaR}_\alpha^\Delta(Z)$  reduces to the conditional expectation of  $-Z$  subject to  $Z < -\text{VaR}_\alpha^\Delta(Z)$ . Then (b) turns into the conditional expectation requirement that

$$E[X_j \mid Z < -\text{VaR}_\alpha^\Delta(Z)] = 0 \text{ for } j = 1, \dots, n. \quad (27)$$

**Detail.** The CVaR subgradient formula in (26) comes from [8, Example 20].  $\square$

Another way of characterizing regression coefficients is available through duality.

**Definition 7** (dual problem in regression). The following optimization problem, in terms of the risk envelope  $\mathcal{Q}$  associated with the deviation measure  $\mathcal{D}$ , will be said to be dual to the optimization  $(\mathcal{P})$  in Definition 4:

$$(\mathcal{P}') \quad \text{maximize } \text{covar}(Q, -Y) \text{ over } Q \in \mathcal{Q} \text{ subject to } \text{covar}(Q, X_j) = 0 \text{ for } j = 1, \dots, n.$$

The duality between problems  $(\mathcal{P})$  and  $(\mathcal{P}')$  emerges from consideration of the Lagrangian function for  $(\mathcal{P}')$ , which takes the form

$$\begin{aligned} L(Q, c_1, \dots, c_n) &= \text{covar}(Q, -Y) + c_1 \text{covar}(Q, X_1) + \dots + c_n \text{covar}(Q, X_n) \\ &= E[Q((EY - Y) - c_1(EX_1 - X_1) - \dots - c_n(EX_n - X_n))]. \end{aligned} \quad (28)$$

**Theorem 5** (regression coefficients as Lagrange multipliers). Suppose  $\mathcal{D}$  is finite everywhere on  $\mathcal{L}^2$ . The conditions in Theorem 4 that characterize the regression coefficients  $c_j$  in terms of an element  $Q$  are equivalent then to having this  $Q$  solve problem  $(\mathcal{P}')$  with  $c_1, \dots, c_n$  as associated Lagrange multipliers, in the sense of those coefficients satisfying:

$$L(Q', c_1, \dots, c_n) \leq L(Q, c_1, \dots, c_n) \text{ for all } Q' \in \mathcal{Q}. \quad (29)$$

They yield in particular the duality relation that

$$[\text{minimum value in } (\mathcal{P})] = [\text{maximum value in } (\mathcal{P}')]. \quad (30)$$

**Proof.** The problem dual to  $(\mathcal{P}')$  in Lagrangian terms consists of minimizing over  $\mathbb{R}^n$  the expression

$$g(c_1, \dots, c_n) = \sup_{Q \in \mathcal{Q}} L(Q, c_1, \dots, c_n).$$

That supremum, however, is  $\mathcal{D}([Y - EY] - c_1[X_1 - EX_1] - \dots - c_n[X_n - EX_n])$  by (10), so  $g$  is the same as the function  $f$  in Proposition 2, which is what gets minimized in  $(\mathcal{P})$ . Hence  $(\mathcal{P})$  and  $(\mathcal{P}')$  do form a dual pair of problems.

Condition (a) of Theorem 4 is equivalent, by Proposition 4, to having the relation  $Q \in \mathcal{Q}_Z$  hold for  $Z = Y - (c_1X_1 + \dots + c_nX_n + d)$ , or equivalently  $Z = (Y - EY) - c_1(X_1 - EX_1) - \dots - c_n(X_n - EX_n)$  (inasmuch as  $EQ = 1$ ). Condition (a) can be identified, therefore, with (29). On the other hand, condition (b) of Theorem 4 is equivalent to having

$$L(Q, c'_1, \dots, c'_n) \geq L(Q, c_1, \dots, c_n) \text{ for all } (c'_1, \dots, c'_n) \in \mathbb{R}^n. \quad (31)$$

These two conditions together mean, in other words, that  $(Q, c_1, \dots, c_n)$  furnishes a saddle point of the Lagrangian  $L$  on  $\mathcal{Q} \times \mathbb{R}^n$ . The duality relation (30) then follows. (For the basic theory of dual optimization problems of convex type, see [4] and [5].)  $\square$

In addition to its intrinsic interest, Theorem 5 provides alternative approaches to computation. Instead of determining the desired regression coefficients  $c_1, \dots, c_n$  by solving the optimization problem  $(\mathcal{P})$  directly, one can solve the dual problem  $(\mathcal{P}')$ , if convenient, and take  $c_1, \dots, c_n$  to be the Lagrange multipliers that come out of that computation. The dual problem will be finite-dimensional, and thus numerically accessible, when the random variables are discrete — with the state space  $\Omega$  being taken as a finite set.

**Example 9** (one-factor regression through duality). *Suppose  $X$  and  $Y$  are discrete random variables with joint probability distribution*

$$P[X = x_k, Y = y_l] = p_{kl} \text{ for } k = 1, \dots, r \text{ and } l = 1, \dots, s,$$

*and we wish to find a best approximation  $Y \approx cX + d$  in the sense of generalized regression with respect to CVaR deviation at a confidence level  $\alpha \in (0, 1)$ . We can identify the states  $\omega \in \Omega$  with the pairs  $(k, l)$  in the specified range. Once we have  $c$ , we will set  $d = \bar{y} - c\bar{x}$ , where  $\bar{x}$  and  $\bar{y}$  are the expectations of  $X$  and  $Y$ . To get  $c$  itself directly, we should solve the problem*

$$\text{minimize } f(c) = \text{CVaR}_\alpha^\Delta(Y - cX) \text{ over } c \in \mathbb{R},$$

*which corresponds to  $(\mathcal{P})$ , but although the function  $f$  is convex and can be represented in various ways conducive to minimization, it is somewhat complicated. Alternatively, we can get  $c$  by a dual approach. For that, we solve the linear programming problem*

$$\text{maximize } \sum_{(k,l) \in \Omega} p_{kl} q_{kl} [\bar{y} - y_l] \text{ over all } q_{kl} \in [0, \alpha^{-1}] \text{ with } \sum_{(k,l) \in \Omega} p_{kl} q_{kl} = 1, \sum_{(k,l) \in \Omega} p_{kl} q_{kl} [\bar{x} - x_k] = 0,$$

*and take as  $c$  the Lagrange multiplier calculated for the final constraint.*

**Detail.** This is immediate from Theorem 5 and the description of the corresponding risk envelope  $\mathcal{Q}$  in Example 1(c). Note that, because of the next-to-final constraint, one could omit  $\bar{y}$  and  $\bar{x}$  from the statement of the maximization problem without affecting the multiplier  $c$  so obtained.  $\square$

Finally, we raise a question that is of interest in understanding the scope of the generalized theory of regression that we have partially been able to outline in this paper. If  $X_1, \dots, X_n$  and  $Y$  are *normally distributed*, would the regression coefficients generated from an arbitrary deviation measure  $\mathcal{D}$  be any different from the classical ones?

This question, by its very nature, only makes sense when posed for deviation measures that are distribution-based, as in Definition 3 and Proposition 1 (but not Example 3).

**Theorem 6** (reduction to classical regression under normality). *Suppose in  $\mathcal{L}^2$  that  $X_1, \dots, X_n$  and  $Y$  are normally distributed r.v.'s, and that  $\mathcal{D}$  is a finite, distribution-based deviation measure. Then the regression coefficients for  $Y$  with respect to  $X_1, \dots, X_n$  and  $\mathcal{D}$  must be the same as the ones obtained from standard deviation.*

**Proof.** To avoid triviality, we can assume that all r.v.'s of the form  $Z = Y - (c_1X_1 + \dots + c_nX_n + d)$  are nonconstant and hence have  $\sigma(Z) > 0$ . Such r.v.'s  $Z$  are all normally distributed in our setting, and the r.v.'s  $Z'$  generated from them by  $Z'(\omega) = [Z(\omega) - EZ]/\sigma(Z)$  thus all have the same distribution, namely the normal distribution with mean 0 and standard deviation 1. Since  $\mathcal{D}$  is distribution-based,  $\mathcal{D}(Z')$  is the same for all such  $Z'$ . Denoting the common value by  $\mathcal{D}_0$ , this number being positive (by axiom D4), we get  $\mathcal{D}(Z) = \sigma(Z)\mathcal{D}_0$ . The minimization of  $\mathcal{D}(Z)$  over the class of r.v.'s  $Z$  in question is therefore the same as the minimization of  $\sigma(Z)$ , and the coefficients yielding the minimum must come out to be the same.  $\square$

## 5 Conclusions

The concept of linear regression has been extended rigorously from the traditional setting to a new formulation in which standard deviation can be replaced by any deviation measure in an axiomatically defined class. Regression coefficients have been shown always to exist, despite asymmetry in the measure and other nonclassical features, and the extra assumptions that assure uniqueness have been identified as well.

Characterizations of the coefficients in terms of risk envelopes and covariance conditions have been worked out. Approaches to computing the coefficients, directly or through a dual problem, have been suggested. Many examples illuminating specific choices of deviation measures have also been supplied. In addition, a question about the role of normality has been answered.

Many questions in the framework of statistics have been left unanswered here, however. For instance, an important open issue is that of determining a test of regression model adequacy (a lack-of-fitness test) for a general deviation measure. Other issues tied to statistical estimation, and connected therefore with the use of generalized deviations in practice, likewise call for future work.

## References

- [1] C. ACERBI, Spectral Measures of Risk: a Coherent Representation of Subjective Risk Aversion, *Journal of Banking and Finance* **26** (2002), 1505–1518.
- [2] P. ARTZNER, F. DELBAEN, J.-M. EBER, D. HEATH, Coherent Measures of Risk, *Mathematical Finance* **9** (1999), 203–227.
- [3] R.T. ROCKAFELLAR, *Integrals which are convex functionals*, Pacific Journal of Math. **24** (1968), 525–540.
- [4] R.T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, 1970.
- [5] R.T. ROCKAFELLAR, *Conjugate Duality and Optimization*, monograph No. 16 in Conference Board of Math. Sciences Series, SIAM Publications, 1974.
- [6] R.T. ROCKAFELLAR, S.P. URYASEV, Optimization of Conditional Value-at-Risk, *Journal of Risk* **2** (2000), 21–42.
- [7] R.T. ROCKAFELLAR, S.P. URYASEV, Conditional Value-at-Risk for General Loss Distributions, *Journal of Banking and Finance* **26** (2002), 1443–1471.
- [8] R.T. ROCKAFELLAR, S.P. URYASEV, M. ZABARANKIN, Deviation Measures in Risk Analysis and Optimization, Research Report 2002-7, Dept. of Industrial and Systems Engineering, University of Florida.